# A Note on Derandomization using the Method of Conditional Expectation

Bader N. Alahmad*

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be our underlying measure space. Let $\phi : \mathbb{R} \to \mathbb{R}$ be a Borel-measurable function (We call such a function Borel). Let $X : \Omega \to \mathbb{R}$ be a Borel-measurable random variable that assumes values in $\{x_i : i \in I\}$ for some $I$. Define

$$\phi(x) = \mathbb{E}(Y \mid X = x) = \sum_{i \in I} \mathbb{E}(Y \mid X = x_i) 1_{\{x = x_i\}}.$$

Then $\phi(x)$ is Borel, since it is the sum of Borel functions $1_{\{x=x_i\}}$. Let

$$\mathbb{E}(Y|X)(\omega) := \phi(X(\omega)) = \sum_{i \in I} \mathbb{E}(Y \mid X = x_i) 1_{\{X = x_i\}}(\omega) \qquad \forall \omega \in \Omega,$$

which can be written as

$$\mathbb{E}(Y|X)(\omega) = \begin{cases} \mathbb{E}(Y \mid X = x_i), & \text{if } X(\omega) = x_i \text{ for some } i \in I \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

But for any event $A \in \mathcal{F}$, we have

$$\begin{aligned} \mathbb{E}(Y \mid A) &= \sum_i y_i \mathbb{P}(Y = y_i \mid A) = \sum_i y_i \frac{\mathbb{P}(\{Y = y_i\} \cap A)}{\mathbb{P}(A)} = \frac{1}{\mathbb{P}(A)} \sum_i y_i \mathbb{E}(1_{\{Y=y_i\}} 1_A) \\ &= \frac{1}{\mathbb{P}(A)} \sum_i \mathbb{E}(y_i 1_{\{Y=y_i\}} 1_A) = \frac{1}{\mathbb{P}(A)} \mathbb{E}\left(1_A \sum_i y_i 1_{\{Y=y_i\}}\right) = \frac{1}{\mathbb{P}(A)} \mathbb{E}(1_A Y) \\ &= \frac{1}{\mathbb{P}(A)} \int_A Y \, \mathrm{d}\mathbb{P}. \end{aligned}$$

Therefore, with $A = \{X = x_i\}$, we get

$$\mathbb{E}(Y|X)(\omega) = \begin{cases} \frac{1}{\mathbb{P}(X=x_i)} \int_{\{X=x_i\}} Y \, \mathrm{d}\mathbb{P}, & \text{if } X(\omega) = x_i \text{ for some } i \in I \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

We note that $\phi(X(\omega))$ is $\sigma(X)$-measurable (not hard to prove.) The tower property of conditional expectation implies that

$$\mathbb{E}(\mathbb{E}(Y \mid X)) = \mathbb{E}(\mathbb{E}(Y \mid X) \mid \mathcal{F}_0) = \mathbb{E}(Y),$$

*Ph.D. student, Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, BC, Canada; bader@ece.ubc.ca.

where $\{\emptyset, \Omega\} = \mathcal{F}_0 \subset \mathcal{F}$ (courser information wins). In particular, it is the last equality that follows by the tower property because if $\mathbb{E}(\mathbb{E}(Y \mid X) \mid \mathcal{F}_0)$ is to be (a version of) the conditional expectation of $\mathbb{E}(Y)$, then for every $A \in \mathcal{F}_0$, the following should hold

$$\int_A \mathbb{E}(\mathbb{E}(Y \mid X) \mid \mathcal{F}_0) \, \mathrm{d}\mathbb{P} = \int_A \mathbb{E}(Y \mid X) \, \mathrm{d}\mathbb{P} = \int_A Y \, \mathrm{d}\mathbb{P},$$

where the last equality follows because $\mathbb{E}(Y \mid X)$ is the conditional expectation of $Y$ conditioned on $\sigma(X)$. Taking $A = \Omega$ we get

$$\int_A \mathbb{E}(\mathbb{E}(Y \mid X) \mid \mathcal{F}_0) \, \mathrm{d}\mathbb{P} = \int_A Y \, \mathrm{d}\mathbb{P} = \int Y \, \mathrm{d}\mathbb{P} = \mathbb{E}(Y).$$

Moreover, from Eqn. (1) and the linearity of expectations we have

$$\mathbb{E}(\mathbb{E}(Y \mid X)) = \sum_{i \in I} \mathbb{E}(Y \mid X = x_i) \mathbb{P}(X = x_i),$$

and thus

$$\mathbb{E}(Y) = \mathbb{E}(\mathbb{E}(Y \mid X)) = \sum_{i \in I} \mathbb{E}(Y \mid X = x_i) \mathbb{P}(X = x_i).$$

Now with this machinery in hand, derandomization using the method of conditional expectations is easy. It applies to the class of algorithms where the random variables constituting the problem are set independently, one at a time. Assume that our problem is to maximize the value of some objective function $W = f(X_1, \ldots, X_n)$, where $f : \mathbb{R}^n \to \mathbb{R}$ is a Borel measurable map. Assume further that we have at our disposal a randomized algorithm that at every step $j$ independently sets $X_j$ to some value, and finally achieves $\mathbb{E}(W) \geqslant \delta \textsc{Opt}$ for some $\delta \leqslant 1$ (thus $W$ and $X_1, \ldots, X_n$ are random variables). Then we can derandomize the algorithm to obtain a sequence of random choices $b_1, \ldots, b_n$ for $X_1, \ldots, X_n$ such that the invariant

$$\delta \textsc{Opt} \leqslant \mathbb{E}(W) \leqslant \mathbb{E}(W \mid X_1 = b_1, \ldots, X_j = b_j)$$

holds at every step $j$, as follows. Consider setting $X_1$. From Eqn. (1) we have

$$\mathbb{E}(W \mid X_1)(\omega) = \sum_{i \in I} \mathbb{E}(W \mid X_1 = x_i) 1_{\{X_1 = x_i\}}(\omega) \qquad \forall \omega \in \Omega. \tag{3}$$

Set $X_1$ to the value $b_1$ that maximizes Eqn. (3), i.e., $b_1$ is the value $x_i$ s.t. $\mathbb{E}(W \mid X_1 = b_1) = \max_{i \in I} \mathbb{E}(W \mid X_1 = x_i)$. This value maximizes $\mathbb{E}(\mathbb{E}(W \mid X_1))$ because

$$\begin{aligned}
\mathbb{E}(\mathbb{E}(W \mid X_1)) &= \sum_{i \in I} \mathbb{E}(W \mid X_1 = x_i) \mathbb{P}(X_1 = x_i) \\
&\leqslant \mathbb{E}(\max_{i \in I} \mathbb{E}(W \mid X_1 = x_i)) \qquad\qquad (4) \\
&= \max_{i \in I} \mathbb{E}(W \mid X_1 = x_i)) \\
&= \mathbb{E}(W \mid X_1 = b_1),
\end{aligned}$$

where inequality (4) follows by monotonicity of expectations:

$$\sum_{i \in I} \mathbb{E}(W \mid X_1 = x_i) 1_{\{X_1 = x_i\}}(\omega) \leqslant \max_{i \in I} \mathbb{E}(W \mid X_1 = x_i) \Longrightarrow$$
$$\mathbb{E}\left(\sum_{i \in I} \mathbb{E}(W \mid X_1 = x_i) 1_{\{X_1 = x_i\}}(\omega)\right) \leqslant \mathbb{E}(\max_{i \in I} \mathbb{E}(W \mid X_1 = x_i))$$

if both expectations are finite. It therefore follows that

$$\delta \textsc{Opt} \leqslant \mathbb{E}(W) = \mathbb{E}(\mathbb{E}(W \mid X_1)) \leqslant \mathbb{E}(W \mid X_1 = b_1).$$

Now consider setting $X_2$, having already set $X_1$ to $b_1$. We have

$$\mathbb{E}(W \mid X_1 = b_1, X_2)(\omega) = \sum_{i \in I} \mathbb{E}(W \mid X_1 = b_1, X_2 = x_i) 1_{\{X_2 = x_i\}}(\omega) \qquad \forall \omega \in \Omega. \tag{5}$$

Set $X_2$ to the value $b_2$ that maximizes Eqn. (5), i.e., $b_2$ is the value $x_i$ s.t. $\mathbb{E}(W \mid X_1 = b_1, X_2 = b_2) = \max_{i \in I} \mathbb{E}(W \mid X_1 = b_1, X_2 = x_i)$. This value maximizes $\mathbb{E}(\mathbb{E}(W \mid X_1 = b_1, X_2))$ because

$$\begin{aligned} \mathbb{E}(\mathbb{E}(W \mid X_1 = b_1, X_2 = x_i)) &= \sum_{i \in I} \mathbb{E}(W \mid X_1 = b_1, X_2 = x_i) \mathbb{P}(X_2 = x_i) \\ &\leqslant \max_{i \in I} \mathbb{E}(W \mid X_1 = b_1, X_2 = x_i) \\ &= \mathbb{E}(W \mid X_1 = b_1, X_2 = b_2), \end{aligned}$$

and it follows that $\mathbb{E}(W \mid X_1 = b_1) = \mathbb{E}(\mathbb{E}(W \mid X_1 = b_1, X_2)) \leqslant \mathbb{E}(W \mid X_1 = b_1, X_2 = b_2)$ (the first equality in this chain follows by the tower property), and thus $\delta \textsc{Opt} \leqslant \mathbb{E}(W) = \mathbb{E}(\mathbb{E}(W \mid X_1)) \leqslant \mathbb{E}(W \mid X_1 = b_1) \leqslant \mathbb{E}(W \mid X_1 = b_1, X_2 = b_2)$. So in general, the invariant $\delta \textsc{Opt} \leqslant \mathbb{E}(W) \leqslant \mathbb{E}(W \mid X_1 = b_1, X_2 = b_2, \ldots, X_j = b_j)$ is maintained at every step $j$ of the algorithm. This explains why this method works !